Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem

Matthieu Raoelison, Wim de Neys

Judgment and Decision Making, Vol. 14, No. 2, March 2019, pp. 170–178

Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem

Matthieu Raoelison* Wim De Neys†

Abstract

The notorious bat-and-ball problem has long been used to demonstrate that people are easily biased by their intuitions. In this paper we test the robustness of biased responding by examining how it is affected by repeated problem presentation. Participants solved 50 standard and control versions of the bat-and-ball problem. To examine the nature of a potential learning effect we adopted a two-response paradigm in which participants have to give a first hunch and can afterwards take the time to deliberate and change their answer. Results showed that both people's first hunches and the responses they gave after deliberation predominantly remained biased from start to finish. But in the rare cases in which participants did learn to correct themselves, they immediately managed to apply the solution strategy and gave a correct hunch on the subsequent problems. We discuss critical methodological and theoretical implications.

Keywords: bat-and-ball, bias, reasoning, decision making, learning

1 Introduction

Decades of research on judgment and decision making have suggested that human thinking is often biased by erroneous intuitions. The pervasiveness of this bias has sometimes led to the characterization of human reasoners as cognitive misers who tend to over-rely on effortless intuitive thinking (Kahneman, 2011). Arguably, one of the most celebrated examples of this phenomenon is the bat-and-ball problem (Frederick, 2005):

> A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?

Intuitively, the answer that readily comes to mind is 10 cents. This is also the answer that the majority of participants presented with the bat-and-ball problem tend to give. However, after some further reflection it will be clear that this answer is incorrect: if the ball costs 10 cents, then the bat – at a dollar more – would cost $1.10, resulting in a total of $1.20. The problem seems to be that people quite naturally parse the $1.10 in $1 and 10 cents, and simply overlook the critical "more than" statement (Kahneman, 2011).

In theory, solving the bat-and-ball problem shouldn't be too hard. It boils down to solving the basic algebraic equation "X + Y = 1.10, Y = 1 + X, Solve for X" – something most educated adults have done at length in their high school math classes (Hoover & Healy, 2017). Nevertheless, the intuitive appeal of the "10 cents" answer seems to have an irresistible pull on people's thinking and leads them astray (Bago, Raoelison & De Neys, 2019; Frederick, 2005).

In this paper we aim to test the pervasiveness of biased responding in the bat-and-ball problem. Our key interest is to examine whether reasoners show any evidence of spontaneous learning when repeatedly solving problems like the bat-and-ball problem. Most classic tasks in the heuristics and biases field – including the bat-and-ball – are "one-shot" problems (Kahneman, 2002, 2011): participants are typically being presented with one single trial. The implicit assumption is that repeated problem presentation might result in a cueing or learning effect and thereby artificially boost performance. However, there is little direct testing of this assumption. Theoretically, a potential spontaneous learning effect would be interesting. If mere repeated problem presentation (i.e., without being explicitly instructed or receiving feedback) helps people to avoid biased (incorrect) responding, this sketches a less bleak picture of human capability. We might be cognitive misers the first time around, but if we manage to spontaneously correct ourselves, this might suggest that our bias and miserliness is less profound than often thought. Obviously, from a more educational point of view this would also point the way towards a cheap and straightforward de-biasing intervention. Testing for an effect of repeated presentation is also important from a methodolog-

This research was supported by a research grant (DIAGNOR, ANR–16–CE28–0010–01) from the Agence Nationale de la Recherche, France.

Raw data can be downloaded from our OSF page: https://osf.io/6aec3/. The preregistered study design can be retrieved from https://osf.io/smqka/register/5771ca429ad5a1020de2872e.

*Paris Descartes University, Sorbonne Paris Cité, UMR 8240 LaPsyDÉ, 46 Rue Saint-Jacques, FR-75005, Paris, France. Email: matthieu.raoelison@ensc.fr.

†Paris Descartes University, Sorbonne Paris Cité, UMR 8240 LaPsyDÉ, France.

ical point of view. Various imaging (e.g., EEG, fMRI) and behavioral methods (e.g., latency analyses) typically require presentation of multiple trials to obtain a reliable signal-to-noise ratio. In this respect it is important to know how stable people's performance is across trials.

Interestingly, some indirect previous evidence suggests that repeated exposure might give rise to a learning effect on the bat-and-ball problem. The bat-and-ball is one of the problems that is featured in the Cognitive Reflection Test (CRT, Frederick, 2005) — a short test that is often administered to measure people's tendency to reflect on their intuitive judgments. This test is frequently included as a predictor in (online) studies (Stewart, Chandler, & Paolacci, 2017; Stagnaro, Pennycook & Rand, 2018; Thomson & Oppenheimer, 2016). It has been shown that performance on the CRT slightly increases with research participation in general and previous CRT exposure in particular (Bialek & Pennycook, 2017; Haigh, 2016; Meyer, Zhou & Frederick, 2018; Stieger & Reips, 2016; Thomson & Oppenheimer, 2016). However, this effect is mainly observed when people are tested with the exact same item content (Chandler, Mueller & Paolacci, 2014; Finucane & Gullion, 2010; Hoover & Healy, 2017). When participants are presented with structurally similar but content-modified items (e.g., "A magazine and a banana cost $2.20 in total. The magazine costs $2 more than the banana. How much does the banana cost?") the association tends to disappear (Chandler et al., 2014; but see also Meyer et al., 2018).

In sum, the little available evidence seems to argue against a strong spontaneous learning effect in the absence of instructions or feedback. However, previous studies were not always specifically designed to test for learning effects and only considered a limited number of trials. Obviously, the fact that people do not learn after having solved one single (or a handful of) problem(s) does not imply that learning is impossible (Frensch & Runger, 2003). In the present study we aim to provide a more conclusive test in a controlled, experimental setting. During a one-hour testing session we presented participants with 50 content-modified bat-and-ball items and additional control problems to examine whether they would eventually be able to spontaneously learn the underlying structure and avoid biased responding.

In addition, we also wanted to examine the nature of a potential learning effect. Recently, a number of studies suggested that reasoners who solve classic reasoning problems correctly, often can do so even when deliberation is experimentally minimized (Bago & De Neys, 2017, 2019; De Neys, 2017; Newman, Gibb & Thompson, 2017). These studies adopt a two-response paradigm (e.g., Thompson, Prowse-Turner & Pennycook, 2011) in which participants initially have to respond to a problem with the first intuitive answer that comes to mind. Immediately afterwards they are given all the time they want to reflect on the problem and give a final answer. To minimize the possibility that reasoners deliberate in the initial response stage, the first response needs to be given under severe time-pressure and/or cognitive load such that participants do not have the time and resources to engage in active deliberation (Bago & De Neys, 2017; Newman et al., 2017). Results indicate that people who give a correct final response after deliberation often already answered correctly at the initial response stage.

We adopted a similar two-response design in the current study. In loose terms, we wanted to ask whether, in case a learning effect occurred, it resulted from a sudden intuitive insight or from more active deliberation (or both). We therefore tracked how repeated presentation affected the initial and final response accuracy, respectively. This also allowed us to test for possible automatization effects. For example, participants might give biased initial and final responses in the beginning of the study, with repeated exposure they might learn to correct themselves when given sufficient time for deliberation, and then finally they might learn to automatize the computations and start producing correct initial responses.

2 Method

2.1 Pre-registration

The study design and sample size was preregistered on the Open Science Framework (https://osf.io/smqka/register/5771ca429ad5a1020de2872e). No specific analyses were preregistered.

2.2 Participants

We recruited 62 participants (38 female, Mean age = 35.5 years, SD = 13.2 years) on Prolific Academic (https://www.prolific.ac).[1] They were paid £5 for their participation. Only native English speakers from Canada, Australia, New Zealand, the United States of America or the United Kingdom were allowed to take part in the study. Among them, 35% (22 participants) reported high school as their highest level of education, while 62% (38 participants) had a higher education degree, and 3% (2 participants aged 14 and 19) reported less than high school as their highest educational level.

2.3 Material

In total 110 items were presented. We first designed 50 variations of the bat-and-ball problem that had the same underlying structure as the original problem but different superficial item content (e.g., "In a building residents have 340 dogs and cats in total. There are 300 more dogs than cats. How many cats are there?"). Each problem specified

[1]Notethat, following our preregistration, we requested only 60 participants on Prolific but two additional participants ended up completing our study. All available data were analyzed.

two types of objects with different quantities instead of prices (e.g., see Bago & De Neys, 2019; Mata et al., 2017; Trouche, 2016). Each of the 50 problems featured unique content with a total amount that was a multiple of ten and ranged from 110 to 650 (see Appendix A).

Each problem was presented with four answer options; the correct response ("5 cents" in the original bat-and-ball), the intuitively cued "heuristic" response ("10 cents" in the original bat-and-ball), and two foil options. Mathematically speaking, the correct equation to solve the standard bat-and-ball problem is: 100 + 2x = 110, instead, people are thought to be intuitively using the "100 + x = 110" equation to determine their response (Kahneman, 2011). We always used the latter equation to determine the "heuristic" answer option, and the former to determine the correct answer option for each problem. Following Bago and De Neys (2019), the two foil options were always the sum of the correct and heuristic answer (e.g., "15 cents" in original bat-and-ball units) and their second greatest common divider (e.g., "1 cent" in original units). For each item, the four response options appeared in a randomly determined order. The following illustrates the full item format:

> In a building residents have 340 dogs and cats in total.
> There are 300 more dogs than cats.
> How many cats are there?
> o 40
> o 60
> o 10
> o 20

One possible cause for a lack of learning effect is that participants simply become bored with the repeated problem presentation and stop paying attention. To avoid that the task would become too repetitive and to verify that participants stayed minimally engaged in the task we also constructed 50 control problems. In the standard bat-and-ball versions the intuitively cued "heuristic" response cues an answer that conflicts with the correct answer. In the "no-conflict" control problems, the heuristic intuition was made to cue the correct response option. This was achieved by deleting the critical relational "more than" statement (e.g., De Neys, Rossi & Houdé, 2013; Travers et al., 2016). With the above example, a control problem version would look as follows:

> In a building residents have 340 dogs and cats in total.
> There are 300 dogs.
> How many cats are there in the building?
> o 40
> o 60
> o 10
> o 20

In this case the intuitively cued "40" answer was also correct. We presented the same four answer options as for a corresponding standard conflict version. We added three words to the control problem question (e.g., "in the building") so that standard "conflict" and control "no-conflict" versions had roughly the same length. Given that the control items can be solved correctly on the basis of mere intuitive reasoning, we expected to see ceiled performance on the control items throughout, if participants are paying minimal attention to the task and refrain from mere random responding.

Finally, in addition to our 50 standard and control items, we also constructed 10 filler problems in which participants simply had to add two quantities. For example,

> In a town, there are 30 Pepsi drinkers and 300 Coke drinkers.
> How many Coke and Pepsi drinkers are there in total?
> o 330
> o 270
> o 90
> o 520

We reasoned that the filler problems would further help to render the task less repetitive and predictable.

In total participants had to solve 110 problems. The problems were grouped into 10 blocks containing each 5 standard problems, 5 control problems, and one filler problem. The filler problem was always presented as the sixth problem in a block. Standard and control problems were presented in a randomized order. Participants could take a short break after completing each block. The content of the standard and control problems in the first and last five blocks was crossed. Items that were presented in their standard version in the first five blocks were presented in their no-conflict version in the last five blocks and vice versa. To avoid familiarity effects, we used the same objects but with a different total quantity for the standard and control version of a problem in the first and last set of 5 blocks.

2.4 Procedure

The experiment was run online on the Qualtrics platform. Participants were specifically instructed that the study would take up to one hour and demanded their full attention throughout. We adopted the two-response procedure from Bago and De Neys (2017, 2019). Participants were instructed they had to provide two consecutive responses for each problem. They were told that we were interested in their very first, initial answer that came to mind and were informed that after selecting their initial response they could reflect on the problem and take as much time as they needed to provide a final answer. To minimize the possibility that participants deliberated during the initial response stage, the initial response

had to be generated within a stringent response deadline and while cognitive resources were burdened with a secondary load task. The deadline for the initial response was set to 5 s, based on the pretesting of Bago and De Neys (2019) who established that this amounted to the time needed to read the problem. The load task was based on the dot memorization task (Miyake, Friedman, Rettinger, Shah & Hegarty, 2001). Before each reasoning problem participants were presented with a complex visual pattern (i.e., 4 crosses in a 3x3 grid) they had to memorize while solving the reasoning problem. After answering the reasoning problem the first time (intuitively), participants were shown four different matrices and had to choose the correct, to-be-memorized pattern (see Appendix B). The load and deadline were applied only during the initial response stage and not during the subsequent final response stage in which participants were allowed to deliberate.

After reading the general instructions participants solved two unrelated practice reasoning problems to familiarize them with the procedure. Next, they solved two practice matrix recall problems (without concurrent reasoning problem). Finally, at the end of the practice, they had to solve the two earlier practice reasoning problems under cognitive load. They were then reminded that there were 110 problems to solve and that they could take a short pause after each block of 11 problems.

Every trial started with a fixation cross shown for 2000 ms. We then presented the first sentence of the problem (e.g., *"In a building residents have 340 dogs and cats in total.")* for 2000 ms. Next, the target pattern for the memorization task was presented for 2000 ms. Afterwards the full problem was presented. At this point participants had 5000 ms to give an answer; after 4000 ms the background of the screen turned yellow to warn participants about the upcoming deadline. If they did not provide an answer before the deadline, they were asked to pay attention to provide an answer within the deadline on subsequent trials.

After the initial response was entered, participants were presented with four matrix patterns from which they had to choose the correct, to-be-memorized pattern. Once they provided their memorization answer, they received feedback as to whether it was correct. If the answer was not correct, they were also asked to pay more attention to memorizing the correct pattern on subsequent trials.

Finally, the same item was presented again, and participants were asked to provide a final response. The presentation order of the response options was always the same in the initial and final response stage but was randomized across trials. Once participants clicked on one of the answer options they were automatically advanced to the next trial.

The color of the answer options was green during the first response, and blue during the final response phase, to visually remind participants of which question they were answering. Therefore, right under the question we also presented a reminder sentence: "Please indicate your very first, intuitive answer." and "Please give your final answer.", respectively, which was also colored as the answer options.

After every block of 11 trials participants were informed that they completed a block and needed to press a button when they were ready to continue with the next block. After the fifth block participants were reminded that they had completed half of the study and were encouraged to try to stay as focused as possible for the remainder of the study.

At the very end of the experiment, participants were shown the standard bat-and-ball problem and were asked whether they had seen it before. We also asked them to enter the solution. Finally, participants completed a page with demographic questions.

2.5 Exclusion criteria

In total, 26 participants reported having seen the bat-and-ball problem before. Seventeen of them (27.4% of all participants) also provided the correct "5 cents" response. Bago and De Neys (2019) excluded these participants to eliminate the possibility that their prior knowledge of the original correct solution would affect the results. Current conclusions were coherent when analyzed with and without application of the exclusion. As one reviewer noted, expressed familiarity might simply be a poor proxy for prior exposure. Following the reviewer's suggestion, all reported results concern the full sample of participants without exclusion.

Participants failed to provide their first answer before the deadline on 117 trials (2% of all trials) and further failed to pick the correct matrix for the load task on 543 trials (9% of remaining trials). Since we could not guarantee that the initial response for these trials did not involve any deliberation, we discarded them and analyzed the 5540 remaining trials (89% of 6200). On average each participant contributed 44.5 (SD = 4.7) standard problem trials and 44.9 (SD = 4.7) control no-conflict trials.

3 Results and discussion

To see whether and what type of learning occurs we looked at how participants changed or did not change their responses throughout the study by performing a direction of change analysis (Bago & De Neys, 2017, 2019). More specifically, on each trial people can give a correct or incorrect response in each of the two response stages. Hence, in theory, this can result in four different types of answer patterns on any single trial ("00", incorrect response in both stages; "11", correct response in both stages; "01", initial incorrect and final correct response; "10", initial correct and final incorrect response). Figure 1 plots the direction of change classification on each of the 50 critical conflict trials for each

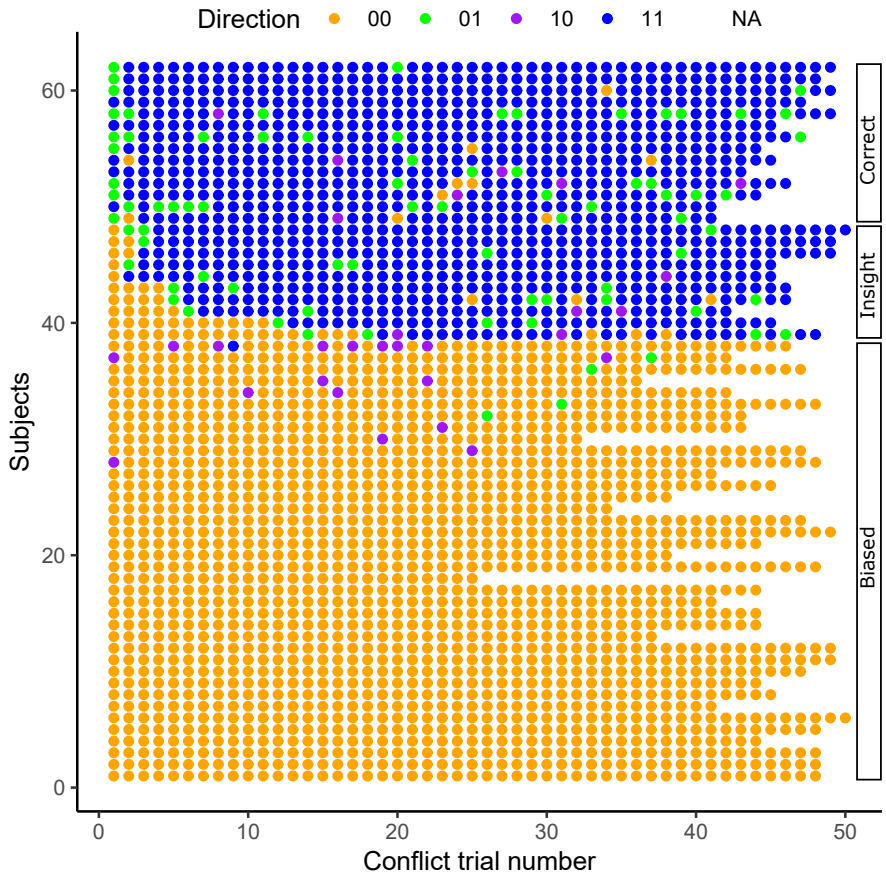

FIGURE 1: Direction of change classification on each conflict trial for each of the 62 subjects. ("00", incorrect initial and final response; "11", correct initial and final response; "01", initial incorrect and final correct response; "10", initial correct and final incorrect response).

individual participant.[2] Eyeballing Figure 1 points to the following descriptive trends:

First, the vast majority of participants (n = 38 out of 62, 61%) predominantly gave 00 responses throughout the study. These participants are labeled as the "biased" group in Figure 1. Both their initial hunch and final deliberate response were typically incorrect from start to finish. Hence, even after very extensive repeated exposure most participants failed to solve the bat-and-ball problem at the end of the study. In and of itself, this seems to confirm the evidence against a strong spontaneous learning effect in previous studies with a more limited number of trials.

Second, 10 participants (out of 62, 16%) started by giving one or more 00 responses but then seemed to achieve insight after a couple of trials and predominantly answered correctly afterwards. These participants are labelled as the "insight" group in Figure 1. Hence, for this small group there is evidence for a spontaneous learning or de-bias effect resulting from mere repeated presentation. It is noteworthy that the insight typically occurred early in the study (i.e., always before trial 20 and often sooner). Moreover, there is also evidence for a fast automatization of the correct solution strategy. In most "insight" cases, the 00 trials are followed by just a single 01 trial after which the subjects predominantly gave 11 responses. Hence, the insight typically occurred during the deliberation phase but this sufficed to solve the subsequent problems correctly during the initial response phase.

Third, 14 participants (out of 62, 23%) started by giving a correct final response on the first trial and typically remained responding correctly till the end. These participants are labelled as the "correct" group in Figure 1. A minority within this group started by giving 11 responses from the beginning (5 out of 14, 36%) but most correct responders started with at least one 01 response after which they quickly gave 11 responses throughout. Hence, as in the "insight" group we see evidence for a fast automatization process. One or two trials in which the correct response is generated after deliberation suffice to generate the correct response as the initial hunch afterwards.

[2]Due to discarding of missed deadline and load trials, not all subjects contributed 50 analyzable trials. Participants in each of the identified groups (see main text) are ranked based on the sum of their total initial and final response accuracy.

3.1 Supplementary analyses

In addition to the 50 conflict problems, participants also solved 50 control, no-conflict problems. Performance on the no-conflict trials was consistently at ceiling throughout the study (average initial response = 96.6%, SD = 18.3%; average final response = 99%, SD = 10%). This can help to argue against a possible general confound. One explanation for the lack of a strong de-bias effect would be that the lengthy nature of the study simply caused most participants to disengage from the task and respond randomly without processing the material. However, the ceiled performance on the control no-conflict problems indicates that participants were paying minimal attention and at least read the problems till the end of the study.

Our main interest in the current study concerned the response accuracy. For completeness, the interested reader can find additional exploratory analyses that look at changes in conflict detection and the type of incorrect response across the study in Appendix C.

4 Discussion

Our results indicate that extensive repeated exposure has overall limited impact on participants' performance on the bat-and-ball problem. For most participants, both their first hunches and the responses they gave after deliberation were predominantly biased from start to finish. Even after solving up to 50 standard problems, heuristic responding typically dominated. At the same time learning was not completely absent. A small group of initially biased reasoners achieved insight after some trials and responded correctly afterwards. Interestingly, there was evidence for a type of fast automatization of the correct solution strategy. Whereas the insight typically occurred during deliberation, the subsequent problems were almost immediately (i.e., after one or two trials) solved correctly during the initial response phase. This same fast automatization was observed among correct responders who had corrected an initial error on the first trial(s). A single instance of correct responding after deliberation often sufficed to give correct initial responses on the subsequent problems.

Although we do not contest that learning was overall rare, we do believe that the "automatization" instances in which it did occur are quite remarkable. In all those cases that people did arrive at the correct response in the deliberation phase, they instantly managed to generate a correct answer during the next initial response stage (i.e., during which deliberation was experimentally minimized). One might wonder what underlies this automatization process: Do people learn to automatize the underlying mathematical equation (e.g., "X + Y = 1.10, Y = 1 + X, Solve for X") or do they rather learn to apply a simpler decision rule (e.g., "take half of the heuristic response"). Our study was not designed to address this issue but note that even if people are simply applying a decision rule this is far from trivial. First, remember that we used content-modified problems with different quantities. Hence, people cannot be merely repeating the previous correct response (e.g., "It's always 5 cents"). Second, blind application of the "take half" rule throughout would lead to errors when solving the control problems. Hence, at the very least participants need to distinguish control and standard problems and recognize that the decision rule is appropriate for the problem at hand. Third, participants need some structural insight to grasp the generalizability of the "take half" principle and realize that the rule is not tied to the specific values in the problem. For example, if the heuristic response is 10 cents and the correct response 5 cents, participants need to properly asses that the correct response on subsequent problems requires halving the heuristic response rather than subtracting "5 cents". In sum, our point is that even the application of a simplified decision rule requires some minimal mathematical insight in the structural relations between the problem values. Although we once again stress that this phenomenon is rare in absolute terms and that most reasoners remain biased throughout, it is noteworthy and should receive more attention from bias researchers in future work.

In this study we presented participants with a total of 110 problems. One might argue that it cannot be excluded that a stronger spontaneous learning effect would occur with a lengthier exposure program in which more problems were presented. We believe this is unlikely. First, the present study took about a full hour. Pushing reasoners even further is very likely to result in fatigue effects that would simply make participants disengage from the task. Second, the study presented both standard and control problems. It has been argued that such a within-subject presentation can help to call participants' attention to the critical problem variable or feature (Kahneman & Fredrick, 2005). In this sense, our task design already created the optimal conditions for spontaneous learning to occur. Third, as Figure 1 indicates, in those cases insight did occur, it typically occurred near the beginning rather than the end of the study, suggesting that adding further trials would have little impact. We therefore believe that it is safe to conclude that simply extending the test session further would not have altered (i.e., improved) the results.

The present study focused on the impact of experimentally manipulated repeated exposure on bat-and-ball problem performance. The overall limited evidence for a spontaneous learning effect fits with previous studies that looked at the impact of naturally occurring variation in exposure to the bat-and ball and/or related problems from the Cognitive Reflection Test (CRT, e.g., Bialek & Pennycook, 2017; Chandler et al., 2014). Note that these previous studies were often also interested in the impact of repeated exposure on the predictive validity of the CRT (e.g., the correlation be-

tween CRT score and, for example, one's susceptibility to heuristics and biases, e.g., Bialek & Pennycook, 2017). The present study did not address this validity issue.

In closing, we would like to stress explicitly that the present study does not imply that it is impossible to learn to avoid biased reasoning. For methodological reasons we were interested in spontaneous learning without instruction or feedback. Obviously, this does not imply that it is futile to try to design intervention or de-bias training programs. For example, previous work already indicated that properly instructing people about the underlying mathematical equation can help to boost performance in the bat-and-ball problem (Hoover & Healey, 2017). Hence, the point is not that people cannot learn to reason correctly. The point is that most people will rarely do this spontaneously when being repeatedly presented with the same problem.

References

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109.

Bago, B., & De Neys, W. (2019). The smart system 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*. http://dx.doi.org/10.1080/13546783.2018.1507949.

Bago, B., Raoelison, M., & De Neys, W. (2019). Second-guess: Testing the specificity of error detection in the bat-and-ball problem. *Acta Psychologica, 193,* 214–228 .

Bialek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods, 50*(5), 1953–1959.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130.

De Neys, W. (Ed.). (2017). *Dual Process Theory 2.0*. London: Routledge.

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, *20*(2), 269–273.

Finucane, M. L., & Gullion, C. M. (2010). Developing a tool for measuring the decision-making competence of older adults. *Psychol Aging*, *25*(2), 271–288.

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*(4), 25–42.

Frensch, P. A., & Rünger, D. (2003). Implicit learning. *Current Directions in Psychological Science*, *12*(1), 13–18.

Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning—in search of a phenomenon. *Thinking & Reasoning*, *21*(4), 383–396.

Haigh, M. (2016). Has the standard cognitive reflection test become a victim of its own success? *Advances in cognitive psychology*, *12*(3), 145.

Hoover, J. D., & Healy, A. F. (2017, Dec). Algebraic reasoning and bat-and-ball problem variants: Solving isomorphic algebra first facilitates problem solving later. *Psychonomic Bulletin & Review*, *24*(6), 1922–1928.

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The doubting system 1: Evidence for automatic substitution sensitivity. *Acta psychologica*, *164*, 56–64.

Kahneman, D. (2002). Maps of bounded rationality: A perspective on intuitive judgment and choice. *Nobel prize lecture*, *8*, 351–401.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. Holyoak & B. Morrison (Eds.), *The cambridge handbook of thinking and reasoning* (pp. 267–293). Cambridge University Press.

Koriat, A. (2017). Can people identify "deceptive" or "misleading" items that tend to produce mostly wrong answers? *Journal of Behavioral Decision Making*, *30*(5), 1066–1077.

Mata, A., Ferreira, M. B., Voss, A., & Kollei, T. (2017). Seeing the conflict: an attentional account of reasoning errors. *Psychonomic Bulletin & Review*, *24*(6), 1980–1986.

Meyer, A., Zhou, E., & Frederick, S. (2018). The non-effects of repeated exposure to the cognitive reflection test. *Judgment and Decision Making*, *13*(3), 246–259.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? a latent-variable analysis. *Journal of experimental psychology: General*, *130*(4), 621.

Newman, I. R., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of experimental psychology. Learning, memory, and cognition*, *43*(7), 1154—1170.

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2018). Performance on the Cognitive Reflection Test is stable across time. *Judgment and Decision Making*, *13*(3), 260–267.

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*, *21*(10), 736–748.

Stieger, S., & Reips, U.-D. (2016). A limitation of the cognitive reflection test: familiarity. *PeerJ*, *4*, e2395.

Szollosi, A., Bago, B., Szaszi, B., & Aczel, B. (2017). Exploring the determinants of confidence in the bat-and-ball problem. *Acta Psychologica*, *180*, 1–7.

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(2), 215–244.

Thompson, V. A., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment & Decision Making*, *11*(1), 99–113.

Travers, E., Rolison, J. J., & Feeney, A. (2016). The time course of conflict on the cognitive reflection test. *Cognition*, *150*, 109–118.

Trouche, E. (2016). *Reasoning as a social competence: an experimental comparison with the intellectualist theories* (Theses, Université de Lyon).

Appendix A. Reasoning material

Material construction clarification. The objects we used to create the problems were derived from 10 categories (animals, fruits, utensils, instruments, groups of people, plants, professions, sports, tools, and vehicles), with every block containing one pair of objects of each category. The first set of five blocks used standard problems derived from the first five categories and control problems from the last five ones. This pattern was reversed for the last set of five blocks. The total quantity ranged between 110 and 650 and was always a multiple of ten (values 200, 300, 400, 500 and 600 were excluded).

The content of the standard and control problems in the first and last five blocks was crossed. Items that were presented in their standard version in the first five blocks were presented in their no-conflict version in the last five blocks and vice versa. To reduce familiarity effects the standard and counterbalanced control versions used the same object content but specified a different total quantity.

Block construction and presentation. Each block had two mini-sets containing either 3 standard and 2 control problems, or 2 standard and 3 control problems, in addition to a filler problem which was always presented in-between mini-sets (i.e. in the sixth position). The order of presentation of the mini-sets was randomized for each participant. Problems within a mini-set were also presented randomly.

The resulting total of 10 main blocks (containing 11 problems each) that we generated were presented in four possible block orders: a. block 1–5, block 6–10, b. block 1–5, block 10–6, c. block 5–1, block 6–10, d. block 5–1, block 10–6.

Example block. Here is an illustration of the problem sequence in a block:

1. In a company there are 150 men and women in total.

   There are 100 more men than women. How many women are there?

   50/5/25/75

2. A science fair has gathered 440 inventors and engineers.

   There are 400 inventors. How many engineers are there in this science fair?

   10/40/20/60

3. In a kitchen there are 260 knives and spoons in total.

   There are 200 more knives than spoons. How many spoons are there?

   90/30/60/15

4. A city has acquired 610 buses and trains in total.

   There are 600 buses. How many trains are there in this city?

   1/15/5/10

5. In a store one can choose between 320 tomatoes and avocados.

   There are 300 more tomatoes than avocados. How many avocados are there?

   30/5/10/20

6. In a town, there are 30 Pepsi drinkers and 300 Coke drinkers.

   How many Coke and Pepsi drinkers are there in total?

   270/90/330/520

7. A national park has 380 roses and lotus flowers in total.

   There are 300 roses. How many lotus flowers are there in this park?

   40/20/80/120

8. In a building residents have 370 dogs and cats in total.

   There are 300 more dogs than cats. How many cats are there?

   35/7/70/105

9. A music store has 210 saxophones and flutes in total.

   There are 200 more saxophones than flutes. How many flutes are there?

   5/10/15/1

10. In a stadium there are 490 volleyball and baseball players.

    There are 400 volleyball. How many baseball players are there in the stadium?

    15/90/135/45

11. In a store there are 550 nails and hammers in total.

    There are 500 nails. How many hammers are there in this store?

    50/75/5/25

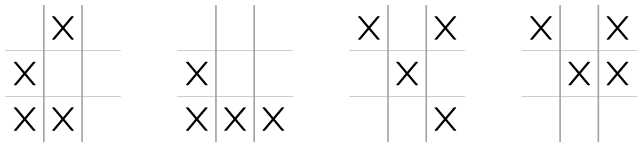

Figure S1. Example of the different response options in the load matrix task: correct answer followed by three distractors.

Full list of problems. A full list of the items used is available in Materials.

Appendix B. Design of load matrices

The patterns were designed using the combination function in R (R Core Team, 2017). Since each 3x3 grid contained four crosses, we used combn(9,4) to list all possible combinations, and placed crosses accordingly. For standard and control problems, we excluded patterns deemed too easy to remember, such as those containing three crosses on a vertical, horizontal, or diagonal line. However, these easier-to-remember patterns could still be used as distractors and were also used as target patterns for the filler problems.

In the response stage, participants were presented with four matrix patterns from which they had to choose the correct, to-be-memorized pattern. The incorrect (distractor) patterns were systematically constructed. The first distractor was the next pattern in our generated R list, meaning that it shared three crosses with the target pattern. The second one was a "complementary" pattern that had no cross in common with the target pattern. Last, the third distractor was the pattern that followed the second distractor in the list. Figure 1 illustrates the construction process of the patterns. All our task material can be retrieved from our OSF page (https://osf.io/6aec3/).

Appendix C. Supplementary analyses

Our key theoretical interest in the present study concerned people's response accuracy. By and large, our accuracy findings showed limited evidence for a strong learning effect. However, it is possible that there might be more subtle learning effects underneath the accuracy surface. For example, previous work has shown that biased responders in the bat-and-ball problem often show some minimal error sensitivity (e.g., De Neys et al., 2013; Gangemi, Bourgeois-Gironde & Mancini, 2015; Johnson, Tubau & De Neys, 2016; Koriat, 2017). Participants seem to doubt their incorrect bat-and-ball responses as reflected in lower response confidence and longer reaction times when responding to standard vs no-conflict control problems. However, not all biased reasoners show this "conflict detection" effect (e.g., Frey, Johnson & De Neys, 2017; Mata, Ferreira, Voss & Kollei, 2017; Szollosi, Bago, Szaszi & Aczel, 2017; Travers, Rolison & Feeney, 2016). It is possible that repeated exposure makes it more likely for biased reasoners to pick up on the conflict and thereby boost the effect. As in previous error or bias detection studies (e.g., Johnson et al., 2016) we therefore calculated a conflict detection index based on participants' response latencies (i.e., response latency for incorrect response on standard problem minus response latency for correct response on control problem.[3]). When contrasting the index across the different trials there was a general descriptive trend towards a decreased effect with repeated presentation This was observed both for the final response (first trial: M = 689 ms, SD = 3697 ms vs last trial: M = 358 ms, SD = 2078 ms) and initial.[4] response (first trial: M = 316 ms, SD = 1180 ms, vs. last trial: M = −23 ms, SD = 833 ms). Hence, if anything there was evidence for a habituation effect. Repeated presentation seemed to make biased reasoners less responsive to the presence of conflict.

Finally, we explored whether the type of incorrect response changed across the study. That is, we always presented participants with four response options: correct response (e.g., "5 cents"), heuristic response (e.g., "10 cents"), and two incorrect foil responses (e.g., "1 cent" and "15 cents"). We simply explored whether there were any changes in the type of incorrect response in the different trials (e.g., were reasoners less attracted by the heuristic response near the end of the study?). However, throughout the study the dominant incorrect response on the standard problems was consistently the heuristic response option [initial response: first trial: M = 94.9%, SD = 22.2%; last trial: M = 94.9%, SD = 22.2%; final response: first trial: M = 100%, SD = 0; last trial: M = 98%, SD = 14.1%]. This consistent high prevalence of one specific type of erroneous response further indicates that the overall low accuracy does not result from a general tendency to disengage from the task and respond randomly.

[3]Reported effects are based on raw RTs but statistical tests for significance have been run on log-transformed RTs. Latency outliers were removed whenever the distance from the average RT in each block x conflict condition was above 3 times the standard deviation. A total of 74 outlying trials (amounting to 1.3% of the total number of trials) were discarded.

[4]We report the initial response data for completeness. Note that initial response latencies have been shown to be a less reliable measure of conflict detection (e.g., Bago & De Neys, 2017; Thompson & Johnson, 2014).